PATENT COOPERATION TREATY

PCT

INTERNATIONAL PRELIMINARY REPORT ON PATENTABILITY
(Chapter II of the Patent Cooperation Treaty)

(PCT Article 36 and Rule 70)

Applicant's or agent's file reference	FOR FURTHER ACTION See Form PCT/IPEA/416	
2020342PC/ko		
International application No.	International filing date *(day/month/year)*	Priority date *(day/month/year)*
PCT/FI 2003/000152	03-03-2003	04-03-2002
International Patent Classification (IPC) or national classification and IPC		
G06N 3/08		
Applicant		
Nokia Corporation et al		

1. This report is the international preliminary examination report, established by this International Preliminary Examining Authority under Article 35 and transmitted to the applicant according to Article 36.
2. This REPORT consists of a total of 4 sheets, including this cover sheet.
3. This report is also accompanied by ANNEXES, comprising:
   a. ☐ *(sent to the applicant and to the International Bureau)* a total of _____ sheets, as follows:
      ☐ sheets of the description, claims and/or drawings which have been amended and are the basis of this report and/or sheets containing rectifications authorized by this Authority (see Rule 70.16 and Section 607 of the Administrative Instructions).
      ☐ sheets which supersede earlier sheets, but which this Authority considers contain an amendment that goes beyond the disclosure in the international application as filed, as indicated in item 4 of Box No. I and the Supplemental Box.
   b. ☐ *(sent to the International Bureau only)* a total of (indicate type and number of electronic carrier(s))
      _____, containing a sequence listing and/or tables related thereto, in computer readable form only, as indicated in the Supplemental Box Relating to Sequence Listing (see Section 802 of the Administrative Instructions).
4. This report contains indications relating to the following items:
   - ☒ Box No. I Basis of the report
   - ☐ Box No. II Priority
   - ☐ Box No. III Non-establishment of opinion with regard to novelty, inventive step and industrial applicability
   - ☐ Box No. IV Lack of unity of invention
   - ☒ Box No. V Reasoned statement under Article 35(2) with regard to novelty, inventive step or industrial applicability; citations and explanations supporting such statement
   - ☐ Box No. VI Certain documents cited
   - ☐ Box No. VII Certain defects in the international application
   - ☐ Box No. VIII Certain observations on the international application

Date of submission of the demand	Date of completion of this report
22-09-2003	28-01-2004
Name and mailing address of the IPEA/SE Patent- och registreringsverket Box 5055 S-102 42 STOCKHOLM Facsimile No. +46 8 667 72 88	Authorized officer Oskar Pihlgren /LR Telephone No. +46 8 782 25 00

INTERNATIONAL PRELIMINARY REPORT ON PATENTABILITY

International Application No.
PCT/FI 2003/000152

Box No. I Basis of the report

1. With regard to the language, this report is based on the international application in the language in which it was filed, unless otherwise indicated under this item.

☐ This report is based on a translation from the original language into the following language _____, which is the language of a translation furnished for the purposes of:

- ☐ international search (under Rules 12.3 and 23.1(b))
- ☐ publication of the international application (under Rule 12.4)
- ☐ international preliminary examination (under Rules 55.2 and/or 55.3)

2. With regard to the elements of the international application, this report is based on *(replacement sheets which have been furnished to the receiving Office in response to an invitation under Article 14 are referred to in this report as "originally filed" and are not annexed to this report):*

☒ the international application as originally filed/furnished

☐ the description:

pages _____ as originally filed/furnished
pages* _____ received by this Authority on _____
pages* _____ received by this Authority on _____

☐ the claims:

pages _____ as originally filed/furnished
pages* _____ as amended (together with any statement) under Article 19
pages* _____ received by this Authority on _____
pages* _____ received by this Authority on _____

☐ the drawings:

pages _____ as originally filed/furnished
pages* _____ received by this Authority on _____
pages* _____ received by this Authority on _____

☐ a sequence listing and/or any related table(s) – see Supplemental Box Relating to Sequence Listing.

3. ☐ The amendments have resulted in the cancellation of:

- ☐ the description, pages _____
- ☐ the claims, Nos. _____
- ☐ the drawings, sheets/figs _____
- ☐ the sequence listing *(specify)*: _____
- ☐ any table(s) related to the sequence listing *(specify)*: _____

4. ☐ This report has been established as if (some of) the amendments annexed to this report and listed below had not been made, since they have been considered to go beyond the disclosure as filed, as indicated in the Supplemental Box (Rule 70.2(c)).

- ☐ the description, pages _____
- ☐ the claims, Nos. _____
- ☐ the drawings, sheets/figs _____
- ☐ the sequence listing *(specify)*: _____
- ☐ any table(s) related to the sequence listing *(specify)*: _____

* *If item 4 applies, some or all of those sheets may be marked "superseded."*

INTERNATIONAL PRELIMINARY REPORT ON PATENTABILITY

International Application No.
PCT/FI 2003/000152

Box No. V **Reasoned statement under Article 35(2) with regard to novelty, inventive step or industrial applicability; citations and explanations supporting such statement**

1. Statement

Novelty (N)	Claims	1-12	YES
	Claims		NO
Inventive step (IS)	Claims	1-12	YES
	Claims		NO
Industrial applicability (IA)	Claims	1-12	YES
	Claims		NO

2. Citations and explanations (Rule 70.7)

Documents cited in the International Search Report:

D1: PATENT ABSTRACTS OF JAPAN, vol.200, no.025, 12 April 2001 (2001-04-12) & JP 2001 229362 A (NIPPON TELEGR & TELEPH CORP) 24 August 2001 (2001-08-24)
abstract

D2: US 6226408 B1

D3: BEZDEK, J.C. et al.: Fuzzy Kohonen clustering networks. IEEE International conference on Fuzzy systems (Cat.no.92CH3073-4), San Diego, CA, USA, 8-12 March 1992. New York 1992, ISBN 0-7083-0236-2, pages 1035-1043

D4: LEE, H.S. et al.: An Investigation into unsupervised clustering techniques. Proceedings of the IEEE SoutheastCon 2000. 'Preparing for the new millenium' (cat.no.00CH37105), Nashville, USA 7-9 April 2000, ISBN 0-7803-6312-4, pages 124-130

D5: Vesanto, J. et al.: Clustering of the Self-Organizing Map. IEEE Trans. Neural Netw.(USA) May 2000, IEEE, USA. ISSN 1045-9227, vol.11 no.3, pages 586-600.

The cited documents represent the general state of the art.
The invention defined in claims 1-12 is not disclosed by any of these documents.
The cited prior art does not give any indication that would lead a person skilled in the art to the claimed method for automatically determining cluster centres.

.../...

Supplemental Box

In case the space in any of the preceding boxes is not sufficient.
Continuation of: Box V

Therefore, the claimed invention is not obvious to a person skilled in the art.

Accordingly, the invention defined in claims 1-12 is novel and is considered to involve an inventive step. The invention is industrially applicable.

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
12 September 2003 (12.09.2003)

PCT

(10) International Publication Number
WO 03/075221 A1

(51) International Patent Classification[7]: G06N 3/08

(21) International Application Number: PCT/FI03/00152

(22) International Filing Date: 3 March 2003 (03.03.2003)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
20020414 4 March 2002 (04.03.2002) FI

(71) Applicant *(for all designated States except US)*: **NOKIA CORPORATION** [FI/FI]; Keilalahdentie 4, FIN-02150 Espoo (FI).

(72) Inventor; and

(75) Inventor/Applicant *(for US only)*: **FLANAGAN, Adrian** [IE/FI]; Vallilantie 36 A 4, FIN-00510 Helsinki (FI).

(74) Agent: **KOLSTER OY AB**; Iso Roobertinkatu 23, P.O.Box 148, FIN-00121 Helsinki (FI).

(81) Designated States *(national)*: AE, AG, AL, AM, AT (utility model), AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ (utility model), CZ, DE (utility model), DE, DK (utility model), DK, DM, DZ, EC, EE (utility model), EE, ES, FI (utility model), FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK (utility model), SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

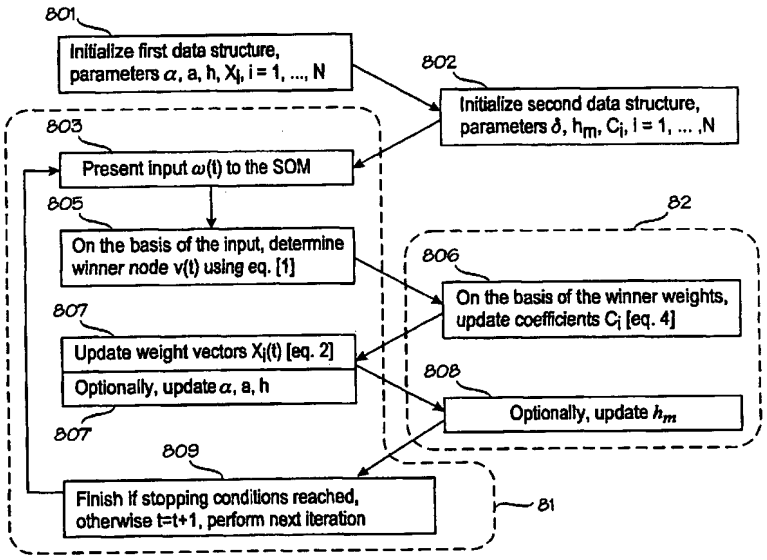
(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PT, SE, SI, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

— *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii)) for the following designations AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH,*

[Continued on next page]

(54) Title: MECHANISM FOR UNSUPERVISED CLUSTERING



(57) Abstract: A computer-implemented method for determining cluster centres in a first data structure, wherein the first data structure comprises a lattice structure of weight vectors that create an approximate representation of a plurality of input data points. The method comprises performing a first iterative process (81) for iteratively updating the weight vectors such that they move toward cluster centres; performing a second iterative process (82) for iteratively updating a second data structure utilizing results of the iterative updating of the first data structure; and determining the weight vectors that correspond to cluster centres of the input data points on the basis of the second data structure.

WO 03/075221 A1

CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, UZ, VC, VN, YU, ZA, ZM, ZW, ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PT, SE, SI, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG)

— *of inventorship (Rule 4.17(iv)) for US only*

Published:

— *with international search report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

MECHANISM FOR UNSUPERVISED CLUSTERING

BACKGROUND OF THE INVENTION

The invention relates to clustering techniques that are generally used to classify input data into groups or clusters without prior knowledge of those clusters. More particularly, the invention relates to methods and apparatus for automatically determining cluster centres. An example of such clustering techniques is a Self-Organizing Map, originally invented by Teuvo Kohonen. The SOM concept is well documented, and a representative example of an SOM application is disclosed in US patent 6 260 036.

The current framework under investigation for describing and analyzing a context has a critical component based on the clustering of data. This clustering is expected to appear at every stage of context computation, from the processing of raw input signals to the determination of a higher order context. Clustering has been well studied over many years and many different approaches to the problem exist. One of the main problems is knowing how many clusters exist in the data. Techniques exist to estimate the number of clusters in a data set, however the methods either require some form of a priori information or assumptions on the data, or they estimate the number of clusters on the basis of an analysis of the data, which may require storing the data, and be computationally demanding. None of these approaches seems entirely suitable for an on-line, unsupervised cluster analysis in a system with limited resources, as would be the case for a context-aware mobile terminal.

Clustering is an important part of any data analysis or information processing problem. The idea is to divide a data set into meaningful subsets so that points in any subset are closely related to each other and not to points in other subsets. The definition of 'related' may be as simple as the distance between the points. Many different approaches and techniques can be applied to achieve this goal. Each approach has its own assumptions and advantages and disadvantages. One of the best-known methods from the partition-based clustering class is the K-means algorithm, which tries to adaptively position K 'centres' that minimize the distance between the input data vectors and the centres. One of its disadvantages is that the number of the K centres must be specified before the clustering is attempted. In the case of an unknown data set this may not always be possible. The algorithm can be run several times with different values of K and the optimum K is chosen on the basis of some

criteria. For an on-line system where the data is not stored, this approach is slow and impractical.

Thus a problem associated with the known clustering techniques is that while it is relatively easy for humans to determine the cluster centres, such a determination is difficult for computers.

BRIEF DESCRIPTION OF THE INVENTION

An object of the invention is to provide a method and an apparatus for implementing the method so as to alleviate the above disadvantages. In other words, the object of the invention is to provide a method for automatically determining cluster centres, such that the method is easily implemented in a computer system.

The object of the invention is achieved by a method and an arrangement which are characterized by what is stated in the independent claims. The preferred embodiments of the invention are disclosed in the dependent claims.

A computer-implemented method according to the invention can be implemented by the following steps:

initializing a first data structure that comprises a lattice structure of weight vectors that create an approximate representation of a plurality of input data points;

performing a first iterative process for iteratively updating the weight vectors such that they move toward cluster centres;

performing a second iterative process for iteratively updating a second data structure utilizing the results of the iterative updating of the first data structure; and

determining, on the basis of the second data structure, the weight vectors that correspond to the cluster centres of the input data points.

A preferred embodiment of the invention is based on the following idea. Self-organizing maps generally use a lattice structure of nodes, and a weight vector is associated with each node. Each data point in the input data is iteratively compared with each weight vector of the lattice, and the node whose weight vector best approximates the data point is chosen as the winner for that data point and iteration. Then the weight vectors associated with each node of the lattice are adjusted. The adjustment made to each node's weight vector is dependent on the winning node through a neighbourhood function. Following

the adjustment of the weight vectors a next iteration step is taken.

As used in this context, the term 'neighbourhood function' is a function of distance on the lattice between the winning node and the node being updated such that the value of the function generally decreases as the distance increases. With normalized SOMs, the value of the function is one for a distance of zero. A common form for the neighbourhood function is Gaussian, but preferred embodiments of the invention make use of neighbourhood functions that are not strictly Gaussian.

In addition to the primary iteration process for updating the SOM, or other clustering mechanism, a second iterative process is run, and the second iterative process gives a numerical value for the lattice nodes such that the numerical value increases if the node's weight vector is positioned at a cluster centre. Then the cluster centres are determined, not on the basis of the weight vectors, but on the basis of the numerical values produced by the second iterative process.

Thus the problem of locating cluster centres reduces to a relatively straightforward problem of locating local maxima in the numerical values produced by the second iterative process.

An advantage of the invention is that it is much easier for machines to locate local maxima in the numerical values than to locate cluster centres in the clustering mechanism wherein the cluster centres are the location in which the density of the weight vectors is highest.

In a preferred embodiment of the invention, the second data structure comprises a coefficient for each of the weight vectors in the lattice structure. Each iteration in the first iterative process comprises selecting a winner weight vector for each of the data points on the basis of a distance measure between the input data point and the weight vector. Each iteration in the second iterative process comprises calculating a next value of each coefficient on the basis of the current value of the coefficient; and a combination of: 1) the current coefficient of the winner weight vector, 2) a second neighbourhood function that approaches zero as the distance on the lattice structure between the weight vector and the winner weight vector increases, and 3) an adjustment factor for adjusting convergence speed between iterations.

The combination referred to above can be a simple multiplication.

If the second neighbourhood function is selected appropriately, such that the second data structure has distinct borders, the step of determining the

weight vectors can be accomplished simply by selecting local maxima in the second data structure.

A preferred version of the second neighbourhood function is not monotonous, but gives negative values at some distances. Also, the second neighbourhood function is preferably made more pronounced over time as the number of prior iterations increases.

Preferably, the first data structure is or comprises a self-organizing map and the input data points represent real-world quantities.

BRIEF DESCRIPTION OF THE DRAWINGS

In the following the invention will be described in greater detail by means of preferred embodiments with reference to the attached drawings, in which

Figure 1 illustrates a self-organizing map (SOM) with six clusters of input data points;

Figure 2 shows a typical form of a neighbourhood function used in an SOM algorithm;

Figure 3 shows a 15 by 15 lattice structure resulting from uniformly distributed input data;

Figure 4 shows a probabilistic map for visualizing the cluster centres in an SOM with six cluster centres;

Figure 5 shows a preferred form of the second neighbourhood function used in the second iterative process according to a preferred embodiment of the invention;

Figure 6 shows a computer pseudocode listing for generating the function shown in Figure 5.

Figure 7 shows a coefficient map that visualizes the data structure used for locating the cluster centres in the SOM; and

Figure 8 is a flow chart illustrating a method according to the invention wherein the method comprises two iterative processes run in tandem;

Figures 9 and 10 show an SOM map and a coefficient map, respectively, with five clusters;

Figures 11 and 12 show an SOM and a coefficient map, respectively, for an exceptional distribution of input data;

Figure 13 shows an example of a neighbourhood function used in an automatic cluster-labelling algorithm;

Figure 14 shows the result of the automatic cluster-labelling algorithm; and

Figure 15 shows how the automatic cluster-labelling algorithm can be integrated with a cluster-determination algorithm according to the invention.

DETAILED DESCRIPTION OF THE INVENTION

A practical example of the invention is disclosed in the context of self-organizing maps. An SOM is a learning algorithm or mechanism from the area of Artificial Neural Networks (ANNs) that find wide application in the area of vector quantization, unsupervised clustering and supervised classification. Reasons for its widespread use include its robustness, even for data sets of very different and even unknown origin, as well as the simplicity of its implementation. The SOM uses a set of weight vectors to form a quantized, topology-preserving mapping, of the input space. The distribution of the weights reflects the probability distribution of the input. The SOM representation is used in clustering applications generally by clustering the weight vectors after training, using for example the K-means algorithm. However the problem of the original K-means algorithm still remains, that is, determining the value of K for the number of centres. In the following, a method based on the SOM algorithm is described which can be used to automatically determine cluster centres in an unsupervised manner. In other words, the number of clusters does not have to be predefined and groups of adjacent SOM weight vectors represent the cluster centres. Unlike the K-means algorithm where each cluster is represented by one centre, in the inventive algorithm the cluster is represented by a set of centres which correspond to weight vectors in the SOM. The algorithm requires few additional computational resources and makes direct use of the information generated during the learning of the SOM. It is already clear why the algorithm can be considered a hybrid of the K-means algorithm and a method based on a probabilistic mixture model. Each cluster is represented by a set of centres, which correspond to a set of weights in the SOM. The SOM weights, in turn, form an approximation of the probability distribution of the input.

From the ANN point of view this may be interesting, as the algorithm uses lateral inhibition between the weight vectors to generate the clusters and a form of Hebbian learning. It is clear that the performance of the clustering depends heavily on the topology-preserving and converging ability of the SOM.

The SOM algorithm

Figure 1 shows a self-organizing map 10. More particularly, Figure 1 shows an online version of SOM. There also exists a "batch SOM" but this requires storing all the input points and going through all of them several times, which is why the online-version is preferred here. Reference numerals 11 generally denote input data points that in most SOM applications represent real-world events or quantities. The SOM algorithm creates an SOM or lattice structure 12 by means of an iterative process that can be summarized as follows. Consider a time sequence of inputs $\omega(t)$, $t = 1, \ldots$, with $\omega \in R^m$ and a probability distribution p_ω. The SOM itself consists of a total of N weight vectors $X \in R^m$ distributed on an n-dimensional lattice. Thus there are two associated dimensions, a dimension m of the input data and the weight vectors, and a dimension n of the lattice. The reason for having a lattice is to be able to define neighbourhood relations between adjacent weights. For example, if each weight k has an associated position vector $i_k \in I^n$ on the lattice, then a distance $d_L(i_j, i_k)$ between weights k and j on the lattice can be defined. The initial values of the weight vectors can be randomly chosen, as the convergence of the algorithm is independent of the initial conditions. At each iteration, a distance $d(\omega(t), X_k)$ between an input $\omega(t)$ at time t and each of the weight vectors X_k is calculated. A winner weight $v(t)$ at time t is then defined as:

$$v(t) = \arg\min_{1 \le k \le N} d(\omega(t), X_k) \qquad [1]$$

For real-valued data, a possible distance measure $d(\cdot,\cdot)$ could be the Euclidean distance. When the winner has been found, each weight vector is updated as:

$$X_k(t+1) = X_k(t) + \alpha(t)h(k, v(t))(\omega(t) - X_k(t)) \qquad [2]$$

where $\alpha(t) \in [0,1]$ is a decreasing function of time which determines the learning rate of the SOM and h is a neighbourhood function which is a function of the distance on the lattice between the winner weight and the updated weight, that is: $h(k, v(t)) = h(d_L(i_{v(t)}, i_k))$.

Figure 2 shows a typical form 21 of a neighbourhood function h used in the SOM algorithm. The neighbourhood function h is positive, having a maximum value of 1 for $d_L = 0$ and decreasing monotonically towards 0 as d_L increases. In practical situations h is often Gaussian, i.e. of the form:

$$h(d_L) = e^{-ad_L^2} \qquad [3]$$

... for some appropriately chosen scaling factor a which may be a function of time. The SOM map is built iteratively. In other words, the steps in equations [1] and [2] are repeated for each t. For uniformly distributed input data and a two-dimensional lattice with 15 × 15 weights, the SOM algorithm gives the result shown in Figure 3.

Figure 3 shows a plot of the weight values on a support [0,1]×[0,1] of the input, where adjacent weights on the lattice are connected by a line segment in the plot. None of the line segments intersect, which indicates that the weights are organized. The weight values are spread uniformly over the input space and approximate the uniform distribution of the input data. To arrive at such a result the learning is divided into two phases. In the first phase, there is a large α and a, and the algorithm is in the self-organizing phase where the weights become organized. The second phase is the convergence phase where the neighbourhood width is reduced to zero as a increases and α approaches zero. The weight vectors converge to the approximation of the probability distribution of the input.

Unsupervised Clustering Based on the SOM

After the description of the basic SOM algorithm, some techniques of determining cluster centres will be disclosed. Consider the SOM shown in Figure 1. Figure 1 shows data points generated from six different normal distributions with different means and variances, and plotted on this there are the weight vectors from the SOM when data from this distribution was used as an input. It is seen that the weight vectors are organized and are more concentrated at the centres of the clusters. Hence the approximation of the probability distribution of the input. Reference numerals 13_1 to 13_6 denote six circles located at the cluster centres. The circles are not normally part of the SOM. While it is easy for humans to determine where the cluster centres are, this task is surprisingly difficult for computers.

First, a probabilistic algorithm for determining the cluster centres will be briefly disclosed. Figure 4 shows a plot of the calculated probability of each weight vector being chosen as a winner during the simulation. The probability for each weight vector is determined by keeping a record of the number of times the weight vector was chosen as a winner and then dividing the number by the total number of iterations in the simulation. The i, j axis indicate the position of each weight on the SOM lattice and the $p(i, j)$ axis shows the probabil-

ity of each of the weights. From this probability distribution a cluster structure is visible where local maxima in the surface correspond to weight vectors positioned at input data cluster centres and local minima form a boundary between the local maximum and hence the clusters. However, the different clusters are not clearly distinct and the surface is not smooth. This roughness of the surface and variations in the peak values of the local maxima may lead to problems when trying to automatically determine the cluster centres using an algorithm which would associate a local maximum of the probability with a cluster centre. Defining a global threshold above which a weight vector would be considered a local maximum and a second global threshold below which a weight vector would be considered a local minimum leads to problems. For example, if in one part of the lattice the probability at a local maximum is below the global threshold to be considered a local maximum hence a cluster centre could lead to erroneous choice of cluster centres. In practise, as in the example shown, this is likely to occur. An alternative would be to define local thresholds for determining a local maximum and a local minimum. However this would be a complicated process requiring an involved computation with no guarantee of a correct result .

A clustering algorithm according to a preferred embodiment of the invention is based on this observation. In effect the algorithm provides a means to smooth the probability surface just described. The use of a neighbourhood function means that the smoothing operation is done locally, emphasising the local maximum as well as the local minimum. The positive elements of the neighbourhood function emphasise the local maximum and the negative components emphasise the local minimum. The result is that all the local maxima consistently reach high values and the local minima consistently reach low values. This allows for the use of a global threshold to identify the maximum and minimum and thus facilitates the use of a computer in the process. Hence instead of using directly the probability of a weight being the winner, a measure somehow related is used in the proposed algorithm which is described as follows.

For each weight i define a scalar coefficient C_i. This coefficient is bounded to the interval [0, 1], and its initial value before training may be quite small. The SOM algorithm is carried out as described earlier. At each iteration the winner weight $v(t)$ is determined as in equation [1] and each SOM weight i

is updated according to equation [2]. At the same time each coefficient C_i is updated as:

$$C_i(t+1) = C_i(t) + C_{v(t)}(t)h_m(d_L)\delta \quad [4]$$

where h_m is the second neighbourhood function. d_L and $v(t)$ are the same terms that were used in the SOM algorithm, that is, d_L is the distance on the lattice between node i and the winner node $v(t)$. δ is a small step value for adjusting convergence speed. δ is somewhat analogous to the α in the SOM algorithm.

Following the update $C_i(t+1)$ is then forced within the interval [0, 1]. For example, if $C_i(t+1) > 1$, it can be set to 1, and if $C_i(t+1) \leq 0$, it can be set to 0.01. Since the update of $C_i(t)$ depends on the value of $C_{v(t)}$, learning is clearly Hebbian.

Figure 5 shows a preferred form 51 of the second neighbourhood function h_m. Actually, Figure 5 shows a time-dependent version of the second neighbourhood function h_m where the function becomes more pronounced as the number of prior iterations increases. In other words, with increasing time, the function h_m achieves negative values sooner (as the distance d_L increases), and the negative values are much more negative than during the earlier iterations.

As can be seen, the preferred form 51 of the second neighbourhood function h_m somewhat resembles the first neighbourhood function h used in the SOM algorithm. Like the first neighbourhood function h, the second neighbourhood function h_m starts at 1 when the distance is zero. Also, h and h_m both approach zero as the distance d_L increases.

However, for some distances, the second neighbourhood function h_m is preferably negative. For instance, in a 10 by 10 lattice, h_m may be negative for distances over 3. The negative value of the second neighbourhood function h_m can be seen as a form of lateral inhibition between the weights. Lateral inhibition is a mathematical model that tries to approximate real biological phenomena. Similar to the h function used in the SOM, weights adjacent to the centre of activity have their coefficients and hence their activity increased, while the activity of weights further away from the centre of activity are inhibited.

This lateral inhibition is rarely if ever used in practical applications, however. In the SOM, the interaction between the weight vectors is defined by

the neighbourhood function h defined by equation [3], which is strictly positive. If h was allowed to be negative at any point, divergence of the weight vectors could result, instead of convergence. In the clustering method proposed here this lateral inhibition is used to determine the cluster centres.

Intuitively it can be seen that if weight i is quite often the winner then C_i will increase along with its neighbours. Similarly, when the winner is i, for its closest neighbours j at a small distance from i on the lattice such that that h_m is positive, the C_j will also increase. At the same time, for j at a large distance from i on the lattice, where h_m is negative, the C_j will decrease. Similarly, if i is not often the winner, its C_i will not increase very much and will be decreased by other winners located at a distance on the lattice. Given the example in Figure 1, it is clear that weights in the inter-cluster regions will have a lower probability of being winners, whereas weights close to the cluster centres will have a higher probability of being winners. Hence it would be expected that the coefficients C_i would be higher for weights positioned in or near the cluster centres and small for positions between the clusters. This would then provide boundaries between the cluster centres. Of course this depends on the fact that the weight vectors X_i reach an organized configuration.

Figure 6 shows an example of a computer pseudocode listing for generating the function shown in Figure 5.

Figure 7 shows a plot 70 of the coefficient values $C(i, j)$ for the weights i, j in the SOM of Figure 1 with $\delta = 0.01$ after 20000 iterations. The second neighbourhood function h_m varied with time and initially did not have a high level of lateral inhibition. Towards the end of the simulation, the level of lateral inhibition was increased. The surface shown in Figure 7 has six distinct and separated elevated regions. As a result of forcing the C_i between 0 and 1, each elevated region corresponds to a set of adjacent weights on the lattice whose coefficients C_i have saturated at or near 1 and whose weights X_i are found at the cluster centre. These regions are surrounded by regions where the coefficients C_i have been driven to small values. Hence it is possible to determine which weights represent the same cluster. To determine which cluster an input vector belongs to, simply find the closest centre and assign the input to the cluster to which the centre belongs. Thus unlike the K-means algorithm where one weight represents the cluster, in the algorithm proposed here a group of weights represents the cluster. The means of classification is the same. Because the coefficients C_i are saturated near 0 or 1, it is a simple task

to determine the cluster centres using a global threshold, the value of which could be set for example at 0.5.

It should be noted that the plot 70 is for visualization purposes only and is not required by computers. Instead, reference number 71 points to an array of current coefficients $C_i(t)$ and reference number 72 to an array of next coefficients $C_i(t+1)$. It is the array 72 of updated coefficients that a computer uses to determine the cluster centres and their locations. An arrow 806 denotes updating of the coefficients that takes place in step 806 of Figure 8 that will be described next.

Figure 8 is a flow chart illustrating a method according to a preferred embodiment of the invention wherein the method comprises two iterative processes 81 and 82 run in tandem. The odd-numbered steps on the left-hand side of Figure 8 relate to the known SOM algorithm. The even-numbered steps on the right-hand side of Figure 8 relate to the inventive algorithm for maintaining and updating the second data structure that is used to determine the cluster centres automatically. In step 801 the SOM algorithm is initialized. The initialization comprises selecting initial values for the α, a, h and randomly initializing the weight vectors X_i. Since Figure 8 shows an online algorithm, the values of the inputs $\omega(t)$ are not known at this stage. Step 802 is a corresponding initialization step for the second data structure. Steps 803, 805, 807 and 809 form the conventional SOM iteration. In step 803, input $\omega(t)$ is presented to the SOM at iteration t. In step 805, the winner weight $v(t)$ is selected according to equation [1]. In step 807, the weight vectors are updated according to equation [2]. In an optional step 807', the variables α that determine learning speed and/or the a of the first neighbourhood function h are updated. In step 809, the iterative loop is repeated until some predetermined stopping criteria are met. For instance, the loop may be set to run a predetermined number of times, or the loop may be interrupted when each succeeding iteration fails to produce a change that exceeds a given threshold.

Steps 806 and 808 relate to the second iterative process 82 for maintaining and updating the second data structure that is used to determine the cluster centres. In step 806, the coefficients C_i are updated on the basis of the winner weights $v(t)$ according to equation [4]. In an optional step 808, parameters for the second neighbourhood function h_m are updated (see Figure 5). According to a preferred feature of the invention, steps 806 and 808 of the second iterative process 82 are interleaved with the steps of the first iterative

process 81. In this way, step 806 utilizes intermediate calculation results of step 805. Similarly, step 808, in which the second neighbourhood function h_m is updated, may utilize data from step 807 that updates the variable α and the first neighbourhood function.

Figures 9 and 10 show an SOM map and a coefficient map, respectively. In this example the number of clusters in the input data distribution was reduced by one and the same SOM algorithm was used. Figure 9 shows the result of both the input and the final configuration of the weight vectors. Figure 10 shows the resulting C_i values. The five cluster centres are once again quite clear. It is remarkable that the same algorithm, without any adjustments, was capable of finding the new number of clusters and the locations of those clusters.

The way the invention works is as follows. In the beginning there is a predefined lattice, which in this case is two-dimensional. Each point of the lattice is given a label, e.g. (2,3), (0,15). This lattice remains fixed and the labelling of the lattice points does not change. In the above examples, the lattice structure is a 15×15 lattice. It is from the lattice that the distances dL used in all neighbourhood functions are determined. For instance, the distance between lattice points (1,3) and (7,8) could be 6, depending on the distance measure we use.

Each lattice point is associated with a weight vector. The dimension of the weight vector is always the same as the dimension of the input data vector. In the examples here the input data has two dimensions. It is the weight vectors that change depending on the input data point and the distance between two points on the lattice, the first point being the lattice point associated with the winner and the second point the lattice point associated with the weight vector to be updated. This distance is not used to update the weight vector directly, but to determine the value of the first neighbourhood function in the update of the weight vector.

Figures 1 and 7 show a two-dimensional plot of the input data points and the weight vectors from the SOM. The plot is in the input space and by chance all the input points were bounded by [-8, 8]. The data points are just the points shown. The weight vectors are plotted at the crossing point of the lines. Another way of looking at this is to draw a line between weight vectors whose associated lattice nodes are the closest nodes on the lattice. The fact that the plot appears as a lattice means that the weights are organized, that is,

the weight vectors associated with adjacent lattice nodes appear in the input data space as adjacent to each other.

The relationship between the coefficients and the SOM lattice is the same as the relationship between the weight vectors and the SOM lattice, except that the dimension of the coefficients is always 1. The relationship is somewhat similar, though not the same as a probability measure, where the probability would be that the weight vector associated with the lattice node with which the coefficient is associated will be chosen as the winner for any given data input. Another interpretation is that the coefficients somehow represent an exaggerated version of the probability distribution of the input data.

In conclusion, we might say that the lattice is a fixed structure. There is one weight vector associated with each lattice node. The weight vector is in the input data space. Similarly, there is one coefficient associated with each lattice node. It is a scalar value and represents an indication of probability, though not the real probability that the weight vector associated with the same lattice node will be chosen as winner for a given input data point.

The dimensionality of the input data and the lattice are not necessarily the same. The input data may have any number of dimensions, such as 5, 10 or 15. In that case the dimensions of the weight vectors would also be 5, 10, or 15. But the lattice could still be two-dimensional. We could also choose a different lattice to begin with (i.e. change the SOM structure) and make it four-dimensional, for example. In this case, if we still choose to have 15 lattice nodes along each axis of the lattice then we would have 15×15×15×15 lattice nodes and associated with each lattice node, a 5, 10 or 15-dimensional weight vector. The examples above use a two-dimensional lattice and a two-dimensional input space merely because it is easier to draw and visualize. In practical implementations one could expect that the input data has more dimensions but the lattice structure could be two-dimensional. The number of lattice nodes along each dimension of the lattice is variable depending on the amount of computational resources available.

Figure 11 shows a mixed clustering example consisting of two very different clusters, namely a first cluster 112 described by a normal distribution and a second cluster 114 described by a uniform distribution along a parabola. The same SOM was used as in the previous two examples. Figure 12 shows a map 120 of coefficients C_{ij} for this example. There are two distinct areas. There is a connected region 122 of high values around three edges of the lat-

tice which correspond to the parabolic distribution cluster 112 in Figure 11, and the disc shape 124 of connected values which corresponds to the normal distribution 114. This example is quite interesting because of the complexity of the parabolic distribution. In the mixture model clustering algorithm, without any prior information it would be difficult to generalize this distribution to a normal distribution. Similarly, a K-means approximation would have difficulty in resolving these two clusters as at some points the distance between two points in different clusters is smaller than the distance between two points in the same cluster. In experiments on this example the best K-means came up with four clusters. This example indicates that the clustering algorithm according to the invention may be very general.

Automatic labelling of cluster centres

A further preferred embodiment of the invention relates to automatic and unsupervised labelling of the clusters. The same notation is used here as above and only notation pertinent to this embodiment will be explained. Consider a set of labels $B = \{1, 2, \ldots, K\}$, which will be used to label the clusters. In practise K should be at least greater than or equal to the expected number of clusters. In the case of no prior knowledge, it may be suitable to let $K = N$, the total number of weights in the SOM, as this imposes a limit on the maximum number of clusters which can be identified.

For each weight i in the SOM, define a vector of coefficients Θ_i as:

$$\Theta i = (\theta_{i,1}, \theta_{i,2}, \ldots \theta_{i,K}) \qquad [5]$$

Each coefficient $\theta_{i,l} \in [0, 1]$ represents a weighting between the SOM node i and the label l. The weight i belongs to cluster l if:

$$l = \underset{1 \le k \le K}{arg\,max}\ \theta_{i,k} \qquad [6]$$

The updating algorithm used on these coefficients to achieve automatic labelling proceeds as follows. At time t SOM weight $v(t)$ is chosen as the winner. The weight and its neighbours are updated as in the normal SOM algorithm. Also the coefficient C_i is updated as well as the coefficients C_j of the neighbours of C_i. The updating of the coefficients C_i and the interpretation of the results form the basis of the main invention, namely the automatic and unsupervised clustering.

In the automatic and unsupervised labelling of the clusters, at the same time t the Θ_i are updated as follows. Define $lv(t)$ as the label of the cluster to which the winner weight $v(t)$ is assigned, thus from equation (6),

$$lv(t) = \underset{1 \le k \le K}{arg\ max}\ \theta_{v(t),k}(t) \qquad [7]$$

For all the weights j, $j = 1, \ldots, N$ the components $\theta_{j,lv(t)}$ are then updated as follows:

$$\theta_{j,lv(t)}(t+1) = \theta_{j,lv(t)}(t) + C_{v(t)}(t+1)\, h_B(d_L)\delta \qquad [8]$$

where once again h_B is a neighbourhood function and preferably has the form shown in Figure 13. The idea is that the neighbours of the winning weight will have their $\theta_{j,lv(t)}$ increased to ensure that they will be classified to the same cluster as the winner $v(t)$, whereas weights further away from the winner weight will have their $\theta_{j,lv(t)}$ decreased to ensure that they will be classified to a different cluster than the winner.

For the weights j where the neighbourhood function $h_B(d_L) > 0$ it is also advantageous to decrease the other coefficients $\theta_{j,lv(t)}$, $k = 1, \ldots K$, $k \neq lv(t)$ in Θ_j as follows:

$$\theta_{j,k}(t+1) = \theta_{j,k}(t) - C_{v(t)}(t+1) h_B(d_L)\delta \qquad [9]$$

This reinforces the labelling of the winner and its neighbours to the cluster label $lv(t)$.

Note that equation [4] uses C_v at iteration t whereas equations [8] and [9] use C_v at iteration t+1. Actually, the C_v coefficient changes so little between iterations that either value can be used, depending on which value is more conveniently available.

Figure 14 shows an example of the result of the combination of the SOM, automatic clustering and automatic labelling of the clusters in the case of an input distribution consisting of five normal distributions. In this case the set of labels was given by $L = \{1, 2, 3, 4, 5, 6, 7, 8\}$. The values of the $\theta_{i,j}$ were randomly initialized. Figure 14 shows an SOM with five distinct regions, one region around each cluster. Each node in the five areas 1 to 5 is assigned a label such that nodes in area 1 have a label of 2, nodes in area 2 have a label of 4, etc. The inter-region areas have a label of 0. These weights had a maximum of Θ_i lower than a threshold value of 0.2 and therefore they are not assigned as centres to any cluster. It is clear that the cluster centres have been properly labelled with the labels $\{2, 4, 5, 6, 8\}$.

Figure 15 shows how the automatic cluster-labelling algorithm can be integrated with the cluster-determination algorithm according to the invention. Figure 15 is a modification of Figure 8, with step 806 followed by step 806' that relates to the automatic cluster-labelling algorithm. In step 806, the cluster labels $lv(t)$ are determined according to equation [7]. Then the $\theta_{j,lv(t)}$ components are updated according to equation [8] and the $\theta_{j,k}$ components are updated according to equation [9]. By placing step 806' inside the second iterative process 82, maximal synergy benefits are obtained, or in other words, the computational overhead is kept to a minimum because step 806' makes used of the winner selection and coefficient determination already performed for the SOM construction and the automatic cluster determination.

Summary

The technique according to the invention allows automatic determination of cluster centres with a minimal amount of information on the data. No explicit, initial estimate of the number of clusters is required. Given the nature of convergence of the SOM, there is no need to know the type of distribution of the clusters either. In this respect the algorithm is very general. However, although explicit initial estimates of the number of clusters are not required, care should be taken to ensure that the SOM lattice contains a number of nodes larger than the expected number of clusters, as well as choosing a non-monotonous neighbourhood function that is negative for large distances and provides a level of lateral inhibition to ensure that the coefficients for the cluster regions stand out more clearly.

The preferred embodiment of the invention, in which the second iterative process is interleaved with the conventional iterative process, requires little computational overhead. Thus this embodiment of the invention is especially suitable for on-line application where human supervision is not available. Initial simulations on artificial data show that the inventive technique is simple and apparently robust and is more easily generalized than most current clustering algorithms. The technique according to the invention can be considered somewhat as a hybrid of the K-means and probabilistic-model-based clustering.

It is readily apparent to a person skilled in the art that, as the technology advances, the inventive concept can be implemented in various ways. The invention and its embodiments are not limited to the examples described

above but may vary within the scope of the claims.

CLAIMS

1. A computer-implemented method for determining cluster centres (13_1 - 13_6) in a first data structure (10), wherein the first data structure comprises a lattice structure (12) of weight vectors that create an approximate representation of a plurality of input data points (11);

the method comprising:

performing a first iterative process (81) for iteratively updating the weight vectors such that they move toward cluster centres (13_1 - 13_6);

performing a second iterative process (82) for iteratively updating a second data structure (70 - 72) utilizing results of the iterative updating of the first data structure; and

determining, on the basis of the second data structure (70 - 72), the weight vectors that correspond to cluster centres of the input data points.

2. A method according to claim 1, wherein each iteration in the first iterative process (81) comprises:

selecting a winner weight vector (v) for each data point on the basis of the distance between the data point and the weight vectors; and

calculating a next value for each weight vector on the basis of the current value of the weight vector and a first neighbourhood function (21, h) of the distance on the lattice structure between the weight vector and the winner weight vector; and

the second data structure (70 - 72) comprises a first coefficient (Ci) for each of the weight vectors in the lattice structure and each iteration in the second iterative process (82) comprises calculating (806) a next value of each first coefficient (Ci) on the basis of:

the current value of the first coefficient; and a combination of:

a first coefficient of the winner weight vector (v),

a second neighbourhood function (51, h_m) of the distance on the lattice structure between the weight vector and the winner weight vector, and

an adjustment factor (δ) for adjusting convergence speed between iterations.

3. A method according to claim 1 or 2, wherein the step of determining the weight vectors that correspond to cluster centres comprises selecting local maxima in the second data structure (70 - 72).

4. A method according to claim 2 or 3, wherein the combination is or comprises multiplication.

5. A method according to any one of claims 2 to 4, wherein the second neighbourhood function (51, h_m) is not monotonous.

6. A method according to any one of claims 2 to 5, wherein the first coefficients are limited to the range [0,1] and the second neighbourhood function (51, h_m) gives negative or positive values, respectively, for some distances.

7. A method according to any one of claims 2 to 6, wherein the second neighbourhood function (51, h_m) depends on the number of prior iterations.

8. A method according to any one of the preceding claims, wherein the input data points (11) represent real-world quantities.

9. A method according to any one of claims 2 to 8, wherein the first data structure (10) is or comprises a self-organizing map.

10. A method according to claim 9, further comprising:

estimating an upper limit K for the number of clusters in the self-organizing map;

defining a coefficient vector $\Theta i = (\theta_{i,1}, \theta_{i,2}, \ldots \theta_{i,K})$ for each weight vector i in the self-organizing map, the coefficient vector comprising K second coefficients $\theta_{i,l}$, each of which represents a weighting between the weight vector i and a label l; and

assigning cluster label l to weight vector i if:

$$l = \underset{1 \le k \le K}{arg\,max}\ \theta_{i,k}.$$

11. A method according to claim 10, wherein each iteration in the second iterative process (82) comprises calculating (806') a next value of each second coefficient on the basis of the current value of the second coefficient and a combination of:

a coefficient of the winner weight vector,

a third neighbourhood function (131, h_B) of distance, and

an adjustment factor (δ) for adjusting convergence speed between iterations.

12. A computer-readable program product comprising a computer program code, wherein executing the computer program code in a computer causes the computer to carry out the steps of the method according to claim 1.

Fig. 1

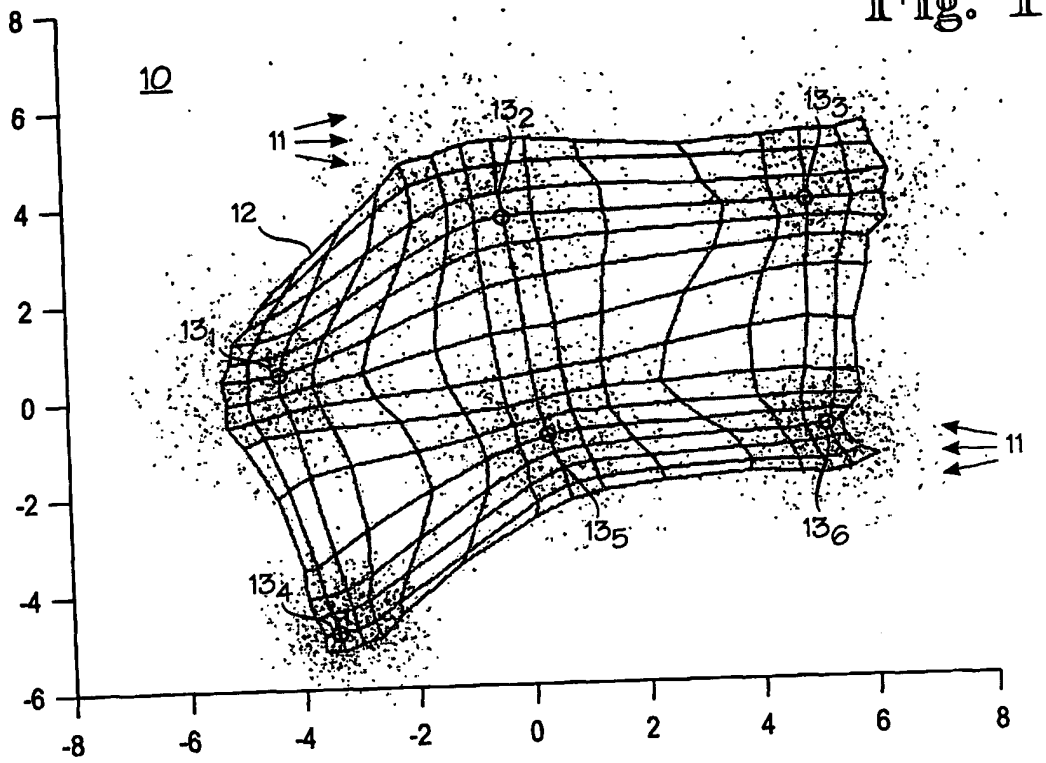


Fig. 2

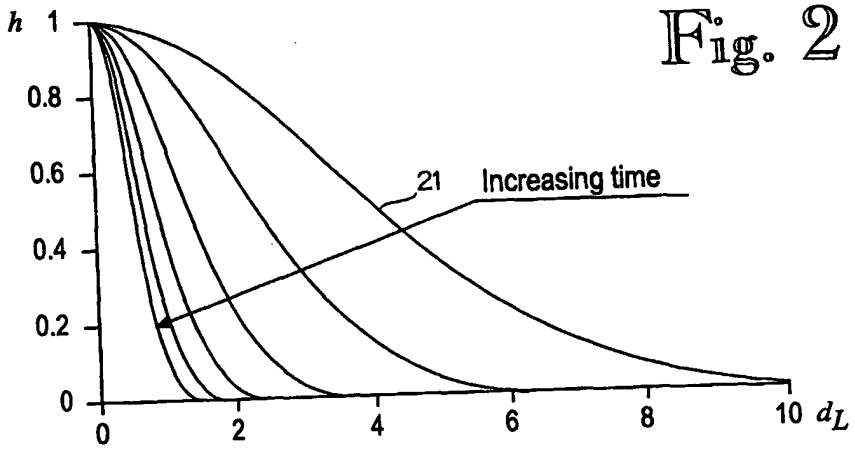


Fig. 3

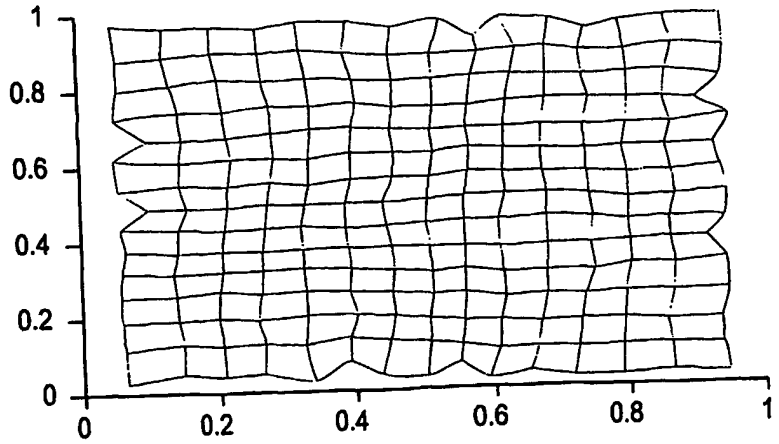


Fig. 4

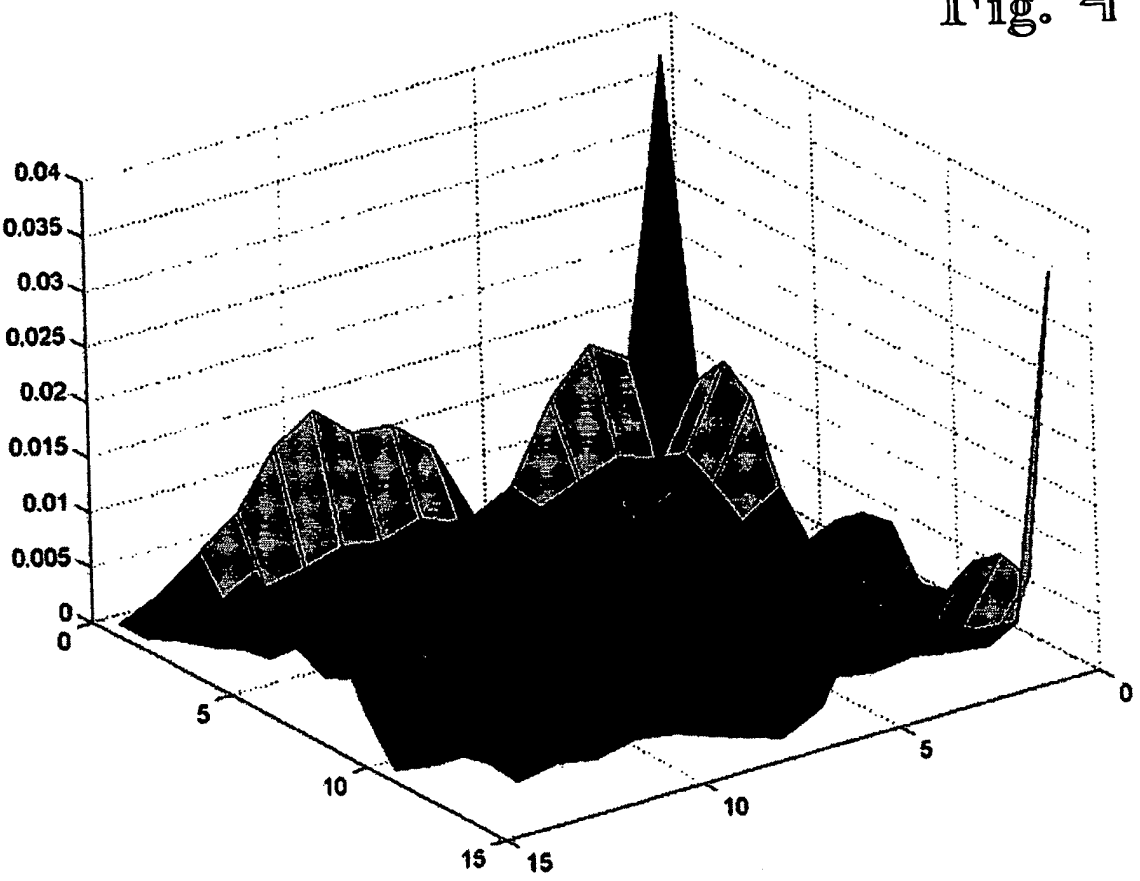

0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
0
5
10
15
15
10
5
0

Fig. 5

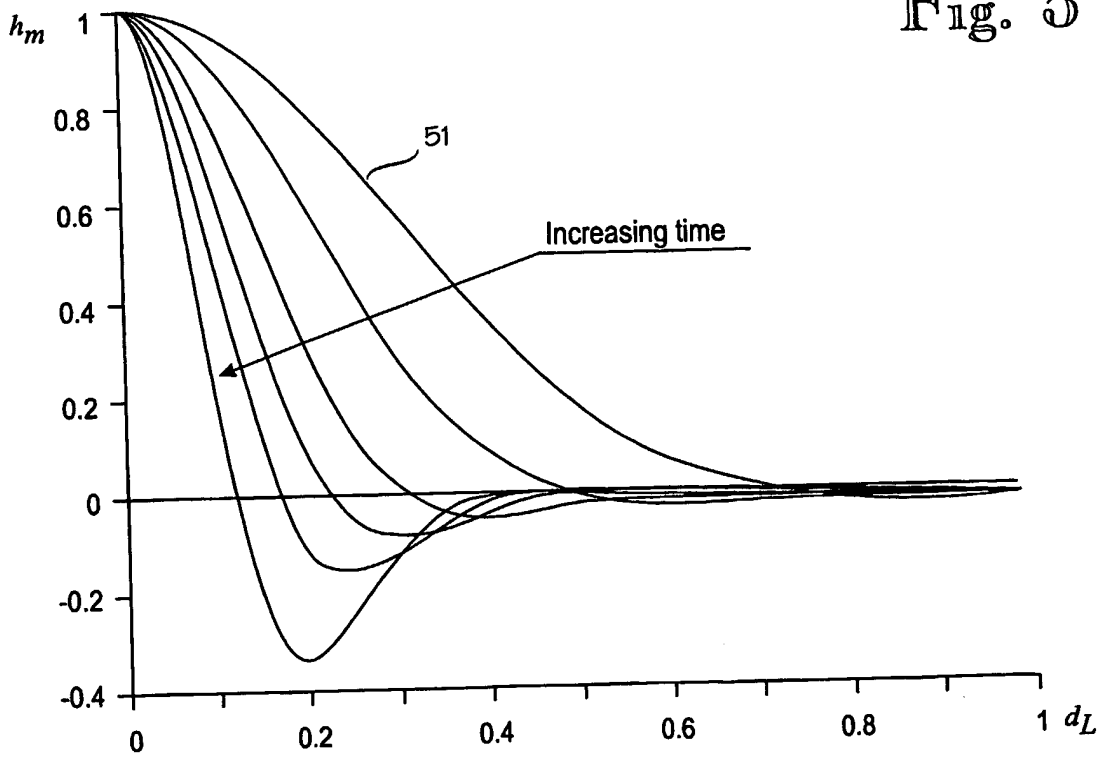

h_m
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
0
0.2
0.4
0.6
0.8
1
d_L
51
Increasing time

Fig. 6

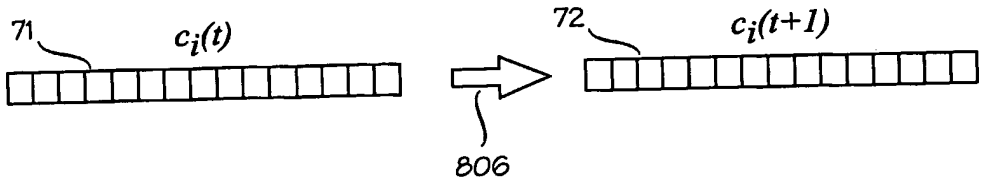
```
x=linspace(0,10,100);

figure

for i=1:1.5:10
    p=0.3+(0.4*(1-(i/10)));
     y=exp(-0.03*x.*x*i).*(p-(0.01*i*x.*x))/p;
     plot(x,y)
     hold on
end
```

60

Fig. 7



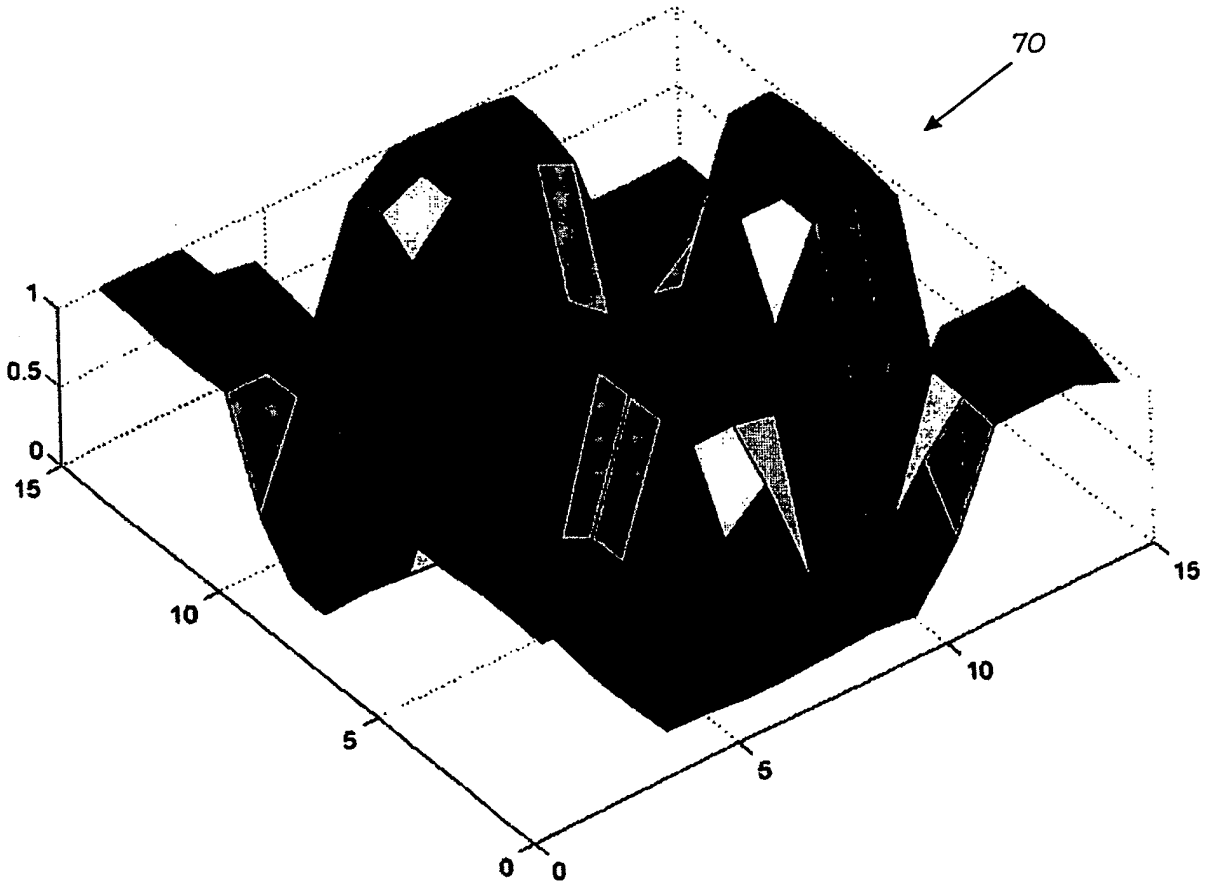


Fig. 8

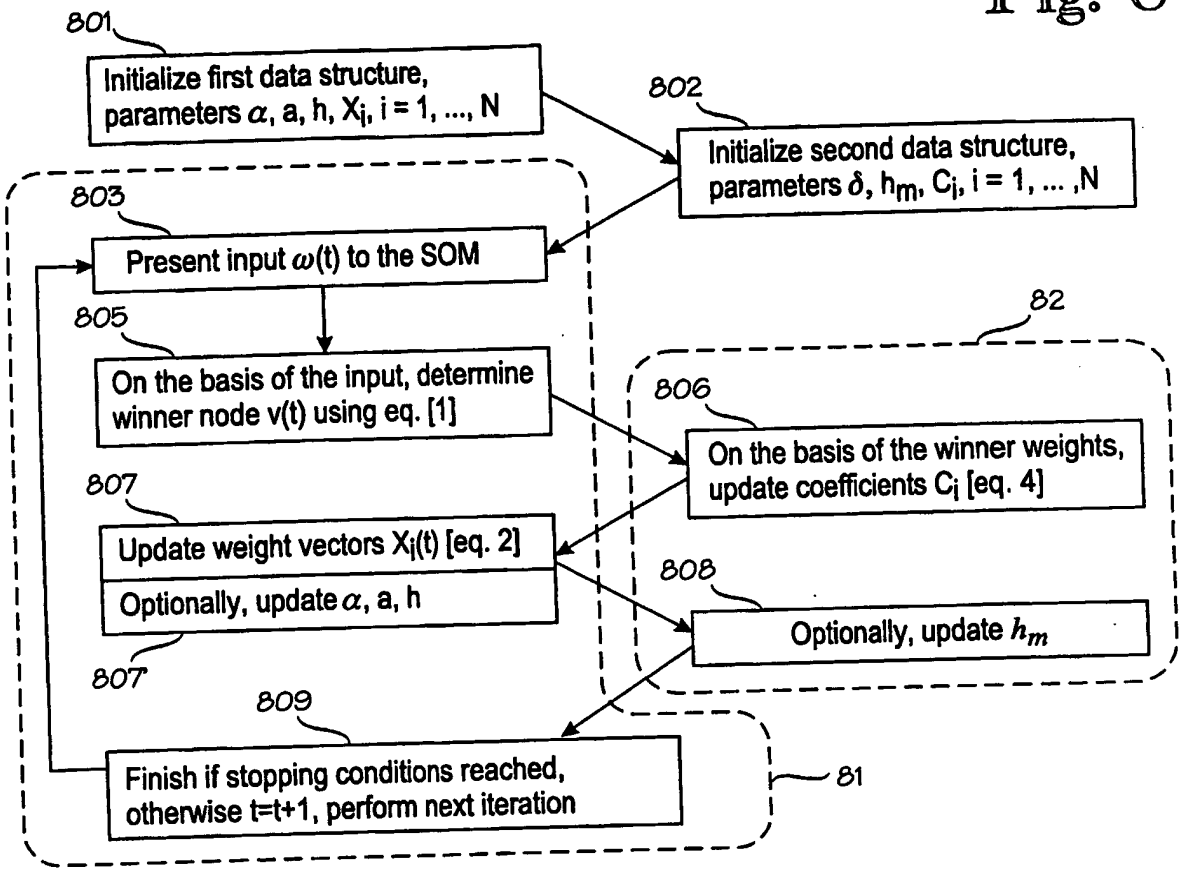

801
Initialize first data structure, parameters α, a, h, Xi, i = 1, ..., N
802
Initialize second data structure, parameters δ, hm, Ci, i = 1, ... ,N
803
Present input ω(t) to the SOM
805
On the basis of the input, determine winner node v(t) using eq. [1]
807
Update weight vectors Xi(t) [eq. 2]
Optionally, update α, a, h
807'
809
Finish if stopping conditions reached, otherwise t=t+1, perform next iteration
81
82
806
On the basis of the winner weights, update coefficients Ci [eq. 4]
808
Optionally, update hm

Fig. 15

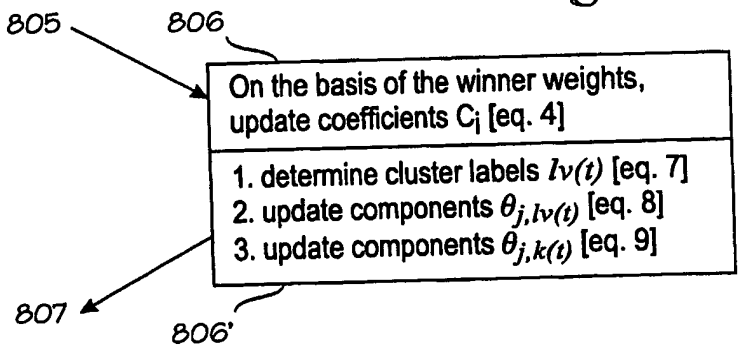

805
806
On the basis of the winner weights, update coefficients Ci [eq. 4]
1. determine cluster labels lv(t) [eq. 7]
2. update components θj,lv(t) [eq. 8]
3. update components θj,k(t) [eq. 9]
807
806'

Fig. 9

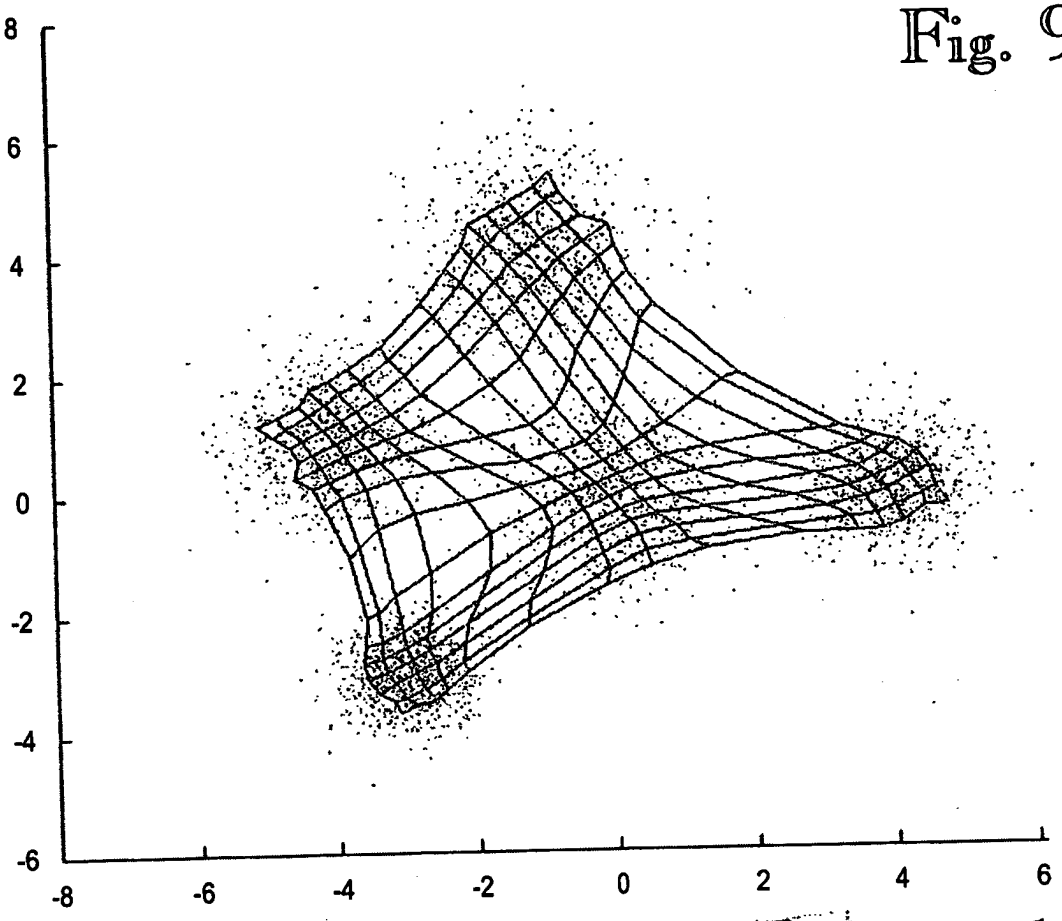


Fig. 10

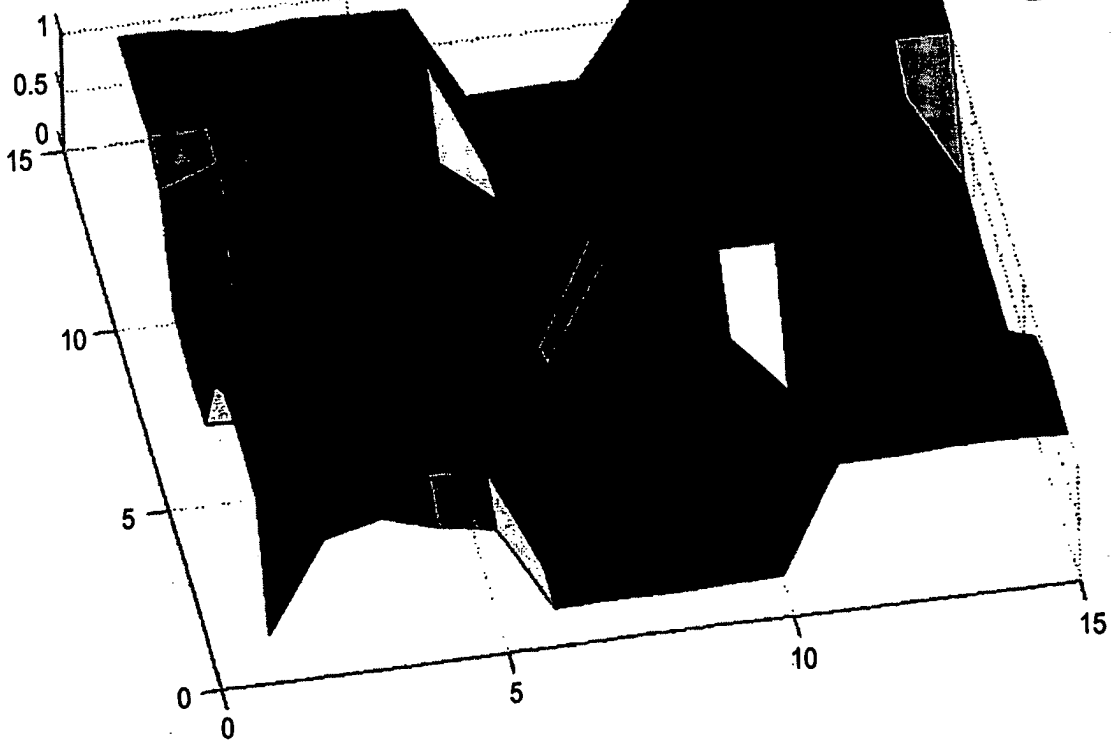


Fig. 11

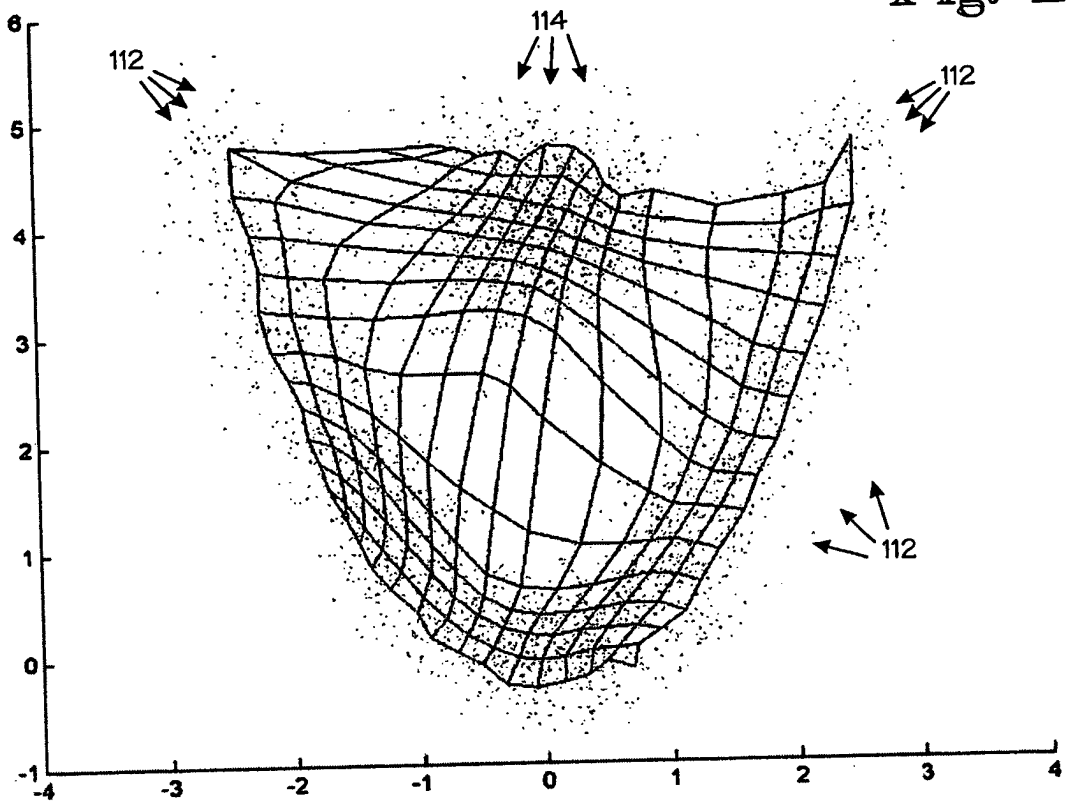

114
112
112
112
6
5
4
3
2
1
0
-1
-4
-3
-2
-1
0
1
2
3
4

Fig. 12

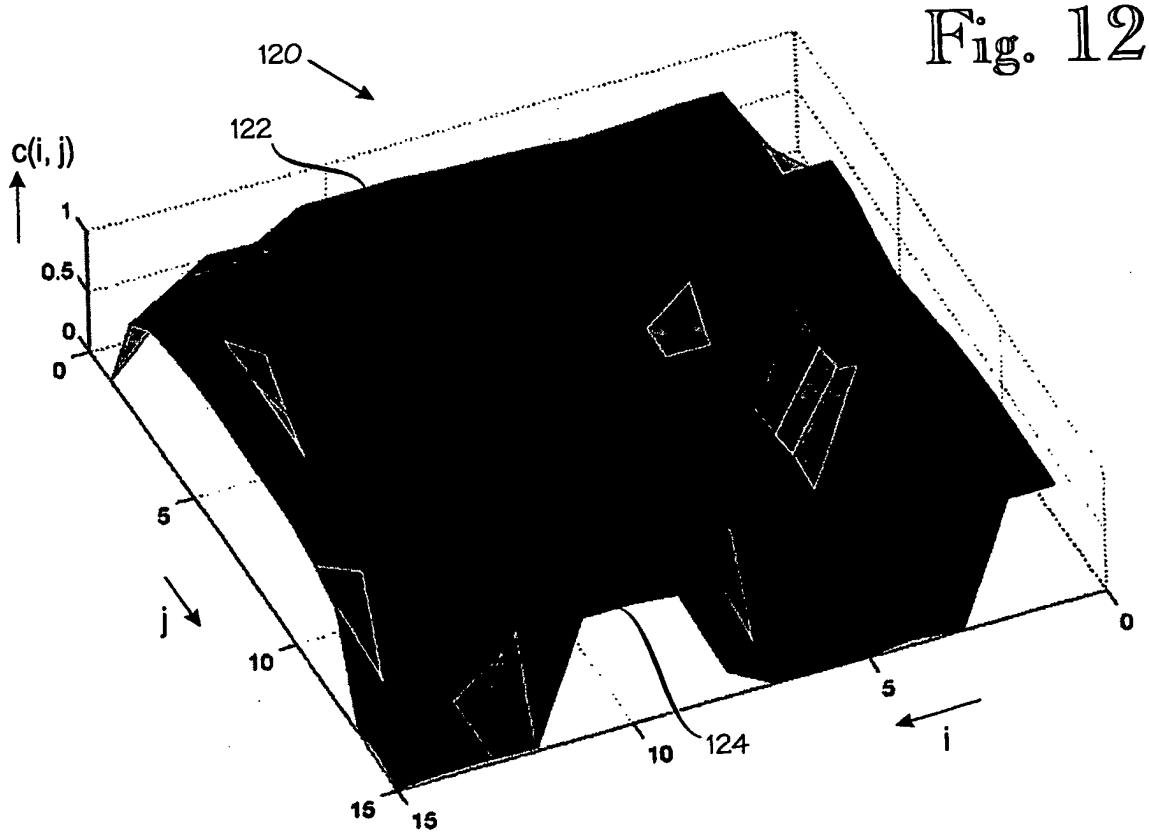

120
122
c(i, j)
1
0.5
0
0
5
j
10
15
15
10
124
5
0
i

Fig. 13

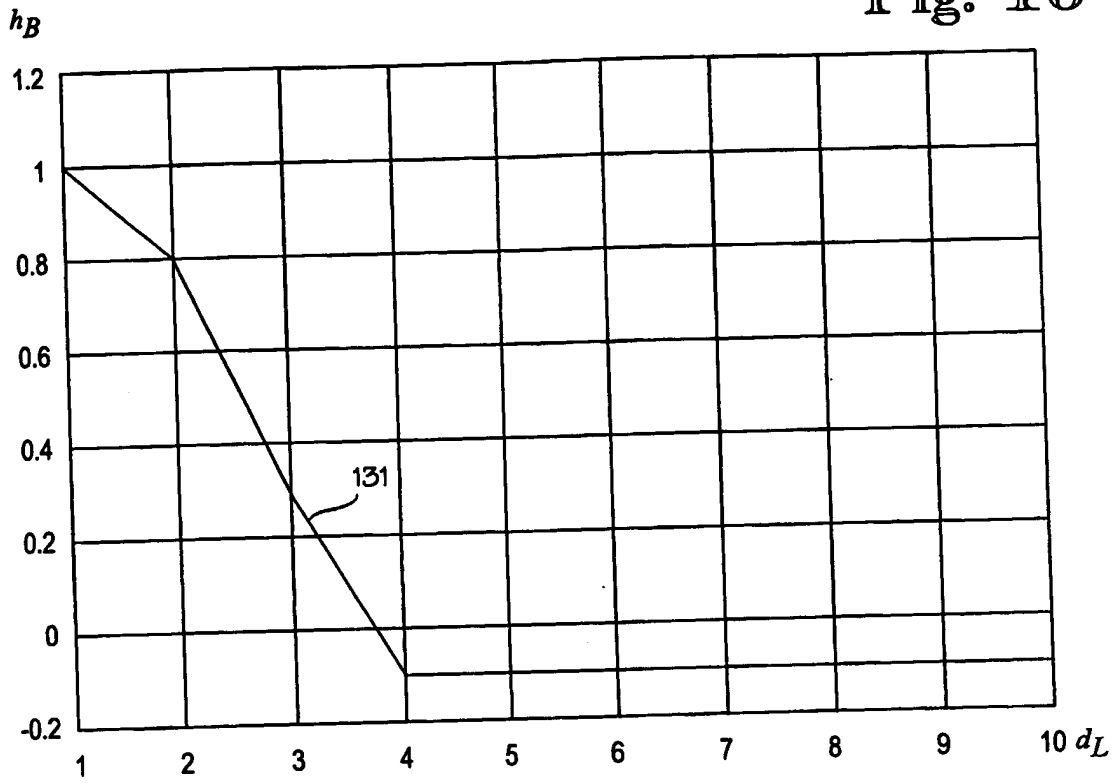
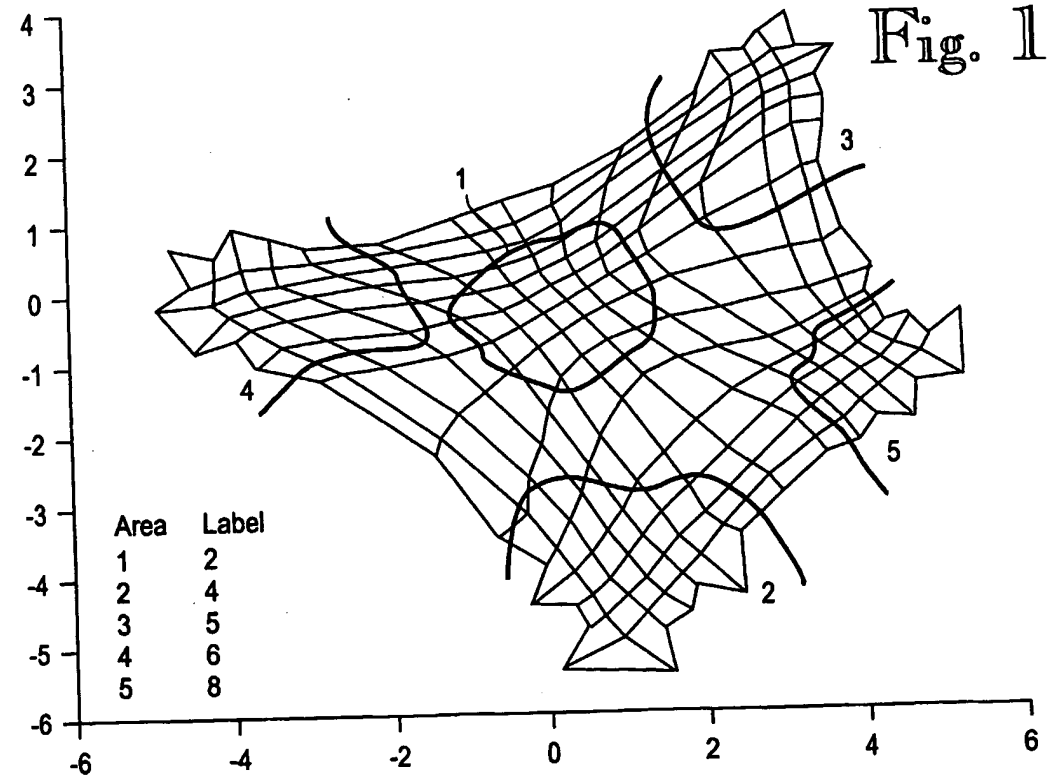

h_B
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
131
1
2
3
4
5
6
7
8
9
10 d_L

Fig. 14


4
3
2
1
0
-1
-2
-3
-4
-5
-6
1
3
4
5
2
Area Label
1 2
2 4
3 5
4 6
5 8
-6
-4
-2
0
2
4
6

INTERNATIONAL SEARCH REPORT

International application No.
PCT/FI[redacted]00152

A. CLASSIFICATION OF SUBJECT MATTER

IPC7: G06N 3/08
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC7: G06N, G06K

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

SE,DK,FI,NO classes as above

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EPO-INTERNAL, WPI DATA, PAJ, INSPEC, COMPENDEX, TDB

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	PATENT ABSTRACTS OF JAPAN vol.200, no.025 12 April 2001 (2001-04-12) & JP 2001 229362 A (NIPPON TELEGR & TELEPH CORP) 24 August 2001 (2001-08-24) abstract --	1-12
A	US 6226408 B1 (SIROSH, J.), 1 May 2001 (01.05.01), column 5, line 50 - column 7, line 43, abstract --	1-12

[X] Further documents are listed in the continuation of Box C. [X] See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance: the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance: the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
22 May 2003	26 -05- 2003
Name and mailing address of the ISA/ Swedish Patent Office Box 5055, S-102 42 STOCKHOLM Facsimile No. +46 8 666 02 86	Authorized officer Jenny Forss /LR Telephone No. +46 8 782 25 00

Form PCT/ISA/210 (second sheet) (July 1998)

INTERNATIONAL SEARCH REPORT

International application No.
PCT/FI [redacted]00152

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	BEZDEK, J.C. et al.: Fuzzy Kohonen clustering networks. IEEE International Conference on Fuzzy systems (Cat.no.92CH3073-4), San Diego, CA, USA, 8-12 March 1992. New York 1992. ISBN 0-7083-0236-2, pages 1035-1043	1-12
A	LEE, H.S. et al.: An investigation into unsupervised clustering technoques. Proceedings of the IEEE SoutheastCon 2000. 'Preparing for the New Millennium' (cat.no.00CH37105), Nashville, USA 7-9 April 2000, ISBN 0-7803-6312-4, pages 124-130	1-12
A	VESANTO, J. et al.: Clustering of the Self-Organizing Map. IEEE Trans. Neural Netw.(USA) May 2000, IEEE, USA. ISSN 1045-9227, vol.11 no.3, pages 586-600	1-12

INTERNATIONAL SEARCH REPORT

Information on pate[redacted]nily members

29/04/03

International application No.
PCT/FI[redacted]/00152

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 6226408 B1	01/05/01	NONE	

PATENT COOPERATION TREATY

From the INTERNATIONAL BUREAU

PCT

NOTICE INFORMING THE APPLICANT OF THE COMMUNICATION OF THE INTERNATIONAL APPLICATION TO THE DESIGNATED OFFICES

(PCT Rule 47.1(c), first sentence)

To:

KOLSTER OY AB
Iso Roobertinkatu 23
P.O.Box 148
FIN-00121 Helsinki
FINLANDE

Date of mailing *(day/month/year)* 12 September 2003 (12.09.03)		
Applicant's or agent's file reference 2020342PC/ko	**IMPORTANT NOTICE**	
International application No. PCT/FI03/00152	International filing date *(day/month/year)* 03 March 2003 (03.03.03)	Priority date *(day/month/year)* 04 March 2002 (04.03.02)
Applicant NOKIA CORPORATION		

1. Notice is hereby given that the International Bureau has **communicated**, as provided in Article 20, the international application to the following designated Offices on the date indicated above as the date of mailing of this notice:

   AU, AZ, BY, CH, CN, CO, DE, DZ, HU, JP, KG, KP, KR, MD, MK, MZ, RU, TM, US

   In accordance with Rule 47.1(c), third sentence, those Offices will accept the present notice as conclusive evidence that the communication of the international application has duly taken place on the date of mailing indicated above and no copy of the international application is required to be furnished by the applicant to the designated Office(s).

2. The following designated Offices have waived the requirement for such a communication at this time:

   AE, AG, AL, AM, AP, AT, BA, BB, BG, BR, BZ, CA, CR, CU, CZ, DK, DM, EA, EC, EE, EP, ES, FI, GB, GD, GE, GH, GM, HR, ID, IL, IN, IS, KE, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MG, MN, MW, MX, NO, NZ, OA, OM, PH, PL, PT, RO, SC, SD, SE, SG, SK, SL, TJ, TN, TR, TT, TZ, UA, UG, UZ, VC, VN, YU, ZA, ZM, ZW

   The communication will be made to those Offices only upon their request. Furthermore, those Offices do not require the applicant to furnish a copy of the international application (Rule 49.1(a-*bis*)).

3. Enclosed with this notice is a copy of the international application as published by the International Bureau on 12 September 2003 (12.09.03) under No. 03/075221

4. **TIME LIMITS for filing a demand for international preliminary examination and for entry into the national phase**

   The applicable time limit for entering the national phase will, **subject to what is said in the following paragraph,** be **30 MONTHS** from the priority date, not only in respect of any elected Office if a demand for international preliminary examination is filed before the expiration of **19 months** from the priority date, but also in respect of any designated Office, in the absence of filing of such demand, where Article 22(1) as modified with effect from 1 April 2002 applies in respect of that designated Office. For further details, see *PCT Gazette* No. 44/2001 of 1 November 2001, pages 19926, 19932 and 19934, as well as the *PCT Newsletter*, October and November 2001 and February 2002 issues.

   In practice, **time limits other than the 30-month time limit** will continue to apply, for various periods of time, in respect of certain designated or elected Offices. For **regular updates on the applicable time limits** (20, 21, 30 or 31 months, or other time limit), Office by Office, refer to the *PCT Gazette*, the *PCT Newsletter* and the *PCT Applicant's Guide*, Volume II, National Chapters, all available from WIPO's Internet site, at http://www.wipo.int/pct/en/index.html.

   For filing a **demand for international preliminary examination**, see the *PCT Applicant's Guide*, Volume I/A, Chapter IX. Only an applicant who is a national or resident of a PCT Contracting State which is bound by Chapter II has the right to file a demand for international preliminary examination (at present, all PCT Contracting States are bound by Chapter II).

   It is the applicant's **sole responsibility** to monitor all these time limits.

The International Bureau of WIPO 34, chemin des Colombettes 1211 Geneva 20, Switzerland	Authorized officer Judith Zahra
Facsimile No.(41-22) 740.14.35	Telephone No.(41-22) 338.91.11

Form PCT/IB/308 (April 2002)

PATENT COOPERATION TREATY

From the INTERNATIONAL BUREAU

PCT

INFORMATION CONCERNING ELECTED OFFICES NOTIFIED OF THEIR ELECTION

(PCT Rule 61.3)

To:

KOLSTER OY AB
Iso Roobertinkatu 23
P.O.Box 148
FIN-00121 Helsinki
Finland

15 -10- 2003

Date of mailing (day/month/year) 08 October 2003 (08.10.03)		
Applicant's or agent's file reference 2020342PC/ko	**IMPORTANT INFORMATION**	
International application No. PCT/FI03/00152	**International filing date** (day/month/year) 03 March 2003 (03.03.03)	**Priority date** (day/month/year) 04 March 2002 (04.03.02)
Applicant NOKIA CORPORATION et al		

1. The applicant is hereby informed that the International Bureau has, according to Article 31(7), notified each of the following Offices of its election:

   EP :AT,BE,BG,CH,CY,CZ,DE,DK,EE,ES,FI,FR,GB,GR,HU,IE,IT,LU,MC,NL,PT,SE,SI,
   SK,TR
   National :BG,CA,CN,DE,GB,IL,JP,KP,KR,MN,NO,PL,RO,RU,SK,US

2. The following Offices have waived the requirement for the notification of their election; the notification will be sent to them by the International Bureau only upon their request:

   AP :GH,GM,KE,LS,MW,MZ,SD,SL,SZ,TZ,UG,ZM,ZW
   EA :AM,AZ,BY,KG,KZ,MD,RU,TJ,TM
   OA :BF,BJ,CF,CG,CI,CM,GA,GN,GQ,GW,ML,MR,NE,SN,TD,TG
   National :AE,AG,AL,AM,AT,AU,AZ,BA,BB,BR,BY,BZ,CH,CO,CR,CU,CZ,DK,DM,DZ,EC,
   EE,ES,FI,GD,GE,GH,GM,HR,HU,ID,IN,IS,KE,KG,KZ,LC,LK,LR,LS,LT,LU,LV,MA,MD,
   MG,MK,MW,MX,MZ,NZ,OM,PH,PT,SC,SD,SE,SG,SL,TJ,TM,TN,TR,TT,TZ,UA,UG,UZ,VC,
   VN,YU,ZA,ZM,ZW

3. The applicant is reminded that he must enter the "national phase" **before the expiration of 30 months from the priority date** before each of the Offices listed above. This must be done by paying the national fee(s) and furnishing , if prescribed, a translation of the international application (Article 39(1)(a)), as well as, where applicable, by furnishing a translation of any annexes of the international preliminary examination report (Article 36(3)(b) and Rule 74.1).

   Some offices have fixed time limits expiring later than the above-mentioned time limit. For detailed information about the applicable time limits and the acts to be performed upon entry into the national phase before a particular Office, see Volume II of the PCT Applicant's Guide.

   The entry into the European regional phase is postponed **until 31 months from the priority date** for all States designated for the purposes of obtaining a European patent.

The International Bureau of WIPO
34, chemin des Colombettes
1211 Geneva 20, Switzerland

Facsimile No. (41-22) 338.87.20

Authorized officer:
El Mostafa MOUSSAID (Fax 338-87-20)

Telephone No. (41-22) 338 9242

5899566

10/506634

Original (for **SUBMISSION**) - printed on 04.03.2003 02:25:00 PM

VIII-4-1	**Declaration: Inventorship (only for the purposes of the designation of the United States of America)** Declaration of inventorship (Rules 4.17(iv) and 51bis.1(a)(iv)) for the purposes of the designation of the United States of America:	I hereby declare that I believe I am the original, first and sole (if only one inventor is listed below) or joint (if more than one inventor is listed below) inventor of the subject matter which is claimed and for which a patent is sought. This declaration is directed to international application No. PCT/FI03/00152(if furnishing declaration pursuant to Rule 26ter) I hereby declare that my residence, mailing address, and citizenship are as stated next to my name. I hereby state that I have reviewed and understand the contents of the above-identified international application, including the claims of said application. I have identified in the request of said application, in compliance with PCT Rule 4.10, any claim to foreign priority, and I have identified below, under the heading "Prior Applications," by application number, country or Member of the World Trade Organization, day, month and year of filing, any application for a patent or inventor's certificate filed in a country other than the United States of America, including any PCT international application designating at least one country other than the United States of America, having a filing date before that of the application on which foreign priority is claimed.
VIII-4-1-1	Prior applications:	20020414, FI, 04 March 2002 (04.03.2002)

		I hereby acknowledge the duty to disclose information that is known by me to be material to patentability as defined by 37 C.F.R. § 1.56, including for continuation-in-part applications, material information which became available between the filing date of the prior application and the PCT international filing date of the continuation-in-part application. I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.
VIII-4-1-1-1	Name:	FLANAGAN, Adrian
VIII-4-1-1-2	Residence: (city and either US State, if applicable, or country)	Helsinki, Finland FIX
VIII-4-1-1-3	Mailing address:	Vallilantie 36 A 14
VIII-4-1-1-4	Citizenship:	~~FI~~ (IRL) IE
VIII-4-1-1-5	Inventor's Signature: (if not contained in the request, or if declaration is corrected or added under Rule 26ter after the filing of the international application. The signature must be that of the inventor, not that of the agent)	Adrian Flanagan
VIII-4-1-1-6	Date: (of signature which is not contained in the request, or of the declaration that is corrected or added under Rule 26ter after the filing of the international application)	18 March 2003

RECORD COPY

2020342PC/ko

PCT REQUEST

Original (for **SUBMISSION**) - printed on 03.03.2003 02:25:00 PM

0	**For receiving Office use only**	
0-1	International Application No.	PCT/FI03/00152
0-2	International Filing Date	03 MAR 2003 (03.03.03)
0-3	Name of receiving Office and "PCT International Application"	The Finnish Patent Office PCT International Application
0-4	**Form - PCT/RO/101 PCT Request**	
0-4-1	Prepared using	PCT-EASY Version 2.92 (updated 01.01.2003)
0-5	**Petition** The undersigned requests that the present international application be processed according to the Patent Cooperation Treaty	
0-6	**Receiving Office (specified by the applicant)**	National Board of Patents and Registration (Finland) (RO/FI)
0-7	**Applicant's or agent's file reference**	2020342PC/ko
I	**Title of invention**	MECHANISM FOR UNSUPERVISED CLUSTERING
II	**Applicant**	
II-1	This person is:	applicant only
II-2	Applicant for	all designated States except US
II-4	Name	NOKIA CORPORATION
II-5	Address:	Keilalahdentie 4 FIN-02150 Espoo Finland
II-6	State of nationality	FI
II-7	State of residence	FI
III-1	**Applicant and/or inventor**	
III-1-1	This person is:	applicant and inventor
III-1-2	Applicant for	US only
III-1-4	Name (LAST, First)	FLANAGAN, Adrian
III-1-5	Address:	Vallilantie 36 A [1]▲▲ 4▲ FIN-00510 Helsinki Finland
III-1-6	State of nationality	[FI]▲▲ IE▲
III-1-7	State of residence	FI

▲RO/FI

▲▲DELETED BY RO/FI

PCT REQUEST

Original (for **SUBMISSION**) - printed on 03.03.2003 02:25:00 PM

IV-1	**Agent or common representative; or address for correspondence** The person identified below is hereby/has been appointed to act on behalf of the applicant(s) before the competent International Authorities as:	agent
IV-1-1	Name	KOLSTER OY AB
IV-1-2	Address:	Iso Roobertinkatu 23 P.O.Box 148 FIN-00121 Helsinki Finland
IV-1-3	Telephone No.	358 9 618 821
IV-1-4	Facsimile No.	+358 9 602 244
V	**Designation of States**	
V-1	Regional Patent (other kinds of protection or treatment, if any, are specified between parentheses after the designation(s) concerned)	AP: GH GM KE LS MW MZ SD SL SZ TZ UG ZM ZW and any other State which is a Contracting State of the Harare Protocol and of the PCT EA: AM AZ BY KG KZ MD RU TJ TM and any other State which is a Contracting State of the Eurasian Patent Convention and of the PCT EP: AT BE BG CH&LI CY CZ DE DK EE ES FI FR GB GR HU IE IT LU MC NL PT SE SI SK TR and any other State which is a Contracting State of the European Patent Convention and of the PCT OA: BF BJ CF CG CI CM GA GN GQ GW ML MR NE SN TD TG and any other State which is a member State of OAPI and a Contracting State of the PCT
V-2	National Patent (other kinds of protection or treatment, if any, are specified between parentheses after the designation(s) concerned)	AE AG AL AM AT (patent and utility model) AU AZ BA BB BG BR BY BZ CA CH&LI CN CO CR CU CZ (patent and utility model) DE (patent and utility model) DK (patent and utility model) DM DZ EC EE (patent and utility model) ES FI (patent and utility model) GB GD GE GH GM HR HU ID IL IN IS JP KE KG KP KR KZ LC LK LR LS LT LU LV MA MD MG MK MN MW MX MZ NO NZ OM PH PL PT RO RU SC SD SE SG SK (patent and utility model) SL TJ TM TN TR TT TZ UA UG US UZ VC VN YU ZA ZM ZW

PCT REQUEST

Original (for **SUBMISSION**) - printed on 03.03.2003 02:25:00 PM

V-5	**Precautionary Designation Statement** In addition to the designations made under items V-1, V-2 and V-3, the applicant also makes under Rule 4.9(b) all designations which would be permitted under the PCT except any designation(s) of the State(s) indicated under item V-6 below. The applicant declares that those additional designations are subject to confirmation and that any designation which is not confirmed before the expiration of 15 months from the priority date is to be regarded as withdrawn by the applicant at the expiration of that time limit.	
V-6	**Exclusion(s) from precautionary designations**	NONE
VI-1	**Priority claim of earlier national application**	
VI-1-1	Filing date	04 March 2002 (04.03.2002)
VI-1-2	Number	20020414
VI-1-3	Country	FI
VI-2	**Priority document request** The receiving Office is requested to prepare and transmit to the International Bureau a certified copy of the earlier application(s) identified above as item(s):	VI-1
VII-1	**International Searching Authority Chosen**	Swedish Patent Office (ISA/SE)
VII-2	**Request to use results of earlier search; reference to that search**	
VII-2-1	Date	12 December 2002 (12.12.2002)
VII-2-2	Number	20020414
VII-2-3	Country (or regional Office)	FI

		Number of declarations	
VIII	**Declarations**		
VIII-1	Declaration as to the identity of the inventor	-	
VIII-2	Declaration as to the applicant's entitlement, as at the international filing date, to apply for and be granted a patent	1	
VIII-3	Declaration as to the applicant's entitlement, as at the international filing date, to claim the priority of the earlier application	-	
VIII-4	Declaration of inventorship (only for the purposes of the designation of the United States of America)	-	
VIII-5	Declaration as to non-prejudicial disclosures or exceptions to lack of novelty	-	

PCT REQUEST

Original (for **SUBMISSION**) - printed on 03.03.2003 02:25:00 PM

VIII-2-1	**Declaration: Entitlement to apply for and be granted a patent** Declaration as to the applicant's entitlement, as at the international filing date, to apply for and be granted a patent (Rules 4.17(ii) and 51bis.1(a)(ii)), in a case where the declaration under Rule 4.17(iv) is not appropriate: Name:	in relation to this international application NOKIA CORPORATION is entitled to apply for and be granted a patent by virtue of the following:
VIII-2-1 (iv)		an assignment from FLANAGAN, Adrian to NOKIA CORPORATION, dated 22 March 2002 (22.03.2002)
VIII-2-1 (ix)	This declaration is made for the purposes of:	all designations except the designation of the United States of America

Original (for **SUBMISSION**) - printed on 03.03.2003 02:25:00 PM

IX	Check list	number of sheets	electronic file(s) attached
IX-1	Request (including declaration sheets)	5	-
IX-2	Description	17	-
IX-3	Claims	3	-
IX-4	Abstract	1	EZABST00.TXT
IX-5	Drawings	7	-
IX-7	TOTAL	33	
	Accompanying items	paper document(s) attached	electronic file(s) attached
IX-8	Fee calculation sheet	✓	-
IX-11	Copy of general power of attorney	reference no. <no.>	-
IX-17	PCT-EASY diskette	-	Diskette
IX-18	Other (specified):	Copy of Official Action	-
IX-19	**Figure of the drawings which should accompany the abstract**	8	
IX-20	**Language of filing of the international application**	English	
X-1	**Signature of applicant, agent or common representative**	Tapio Valkeiskangas	
X-1-1	Name	KOLSTER OY AB	

FOR RECEIVING OFFICE USE ONLY

10-1	**Date of actual receipt of the purported international application**	03 MAR 2003 (03.03.03)
10-2	**Drawings:**	
10-2-1	Received	
10-2-2	Not received	
10-3	**Corrected date of actual receipt due to later but timely received papers or drawings completing the purported international application**	
10-4	**Date of timely receipt of the required corrections under PCT Article 11(2)**	
10-5	**International Searching Authority**	ISA/SE
10-6	**Transmittal of search copy delayed until search fee is paid**	

FOR INTERNATIONAL BUREAU USE ONLY

11-1	**Date of receipt of the record copy by the International Bureau**	